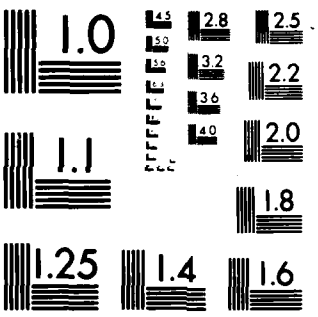
AD-A129 529 CONCURRENT UPDATES AND RETRIEVAL IN DISTRIBUTED DATABASE SYSTEMS(U) CALIFORNIA UNIV BERKELEY ELECTRONICS RESEARCH LAB M R STONEBRAKER ET AL. JAN 83 1/1

UNCLASSIFIED AFOSR-TR-83-0512 AFOSR-78-3596 F/G 5/2 NL

END
DATE FILMED
#7 83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

FINAL REPORT

CONCURRENT UPDATES AND RETRIEVAL IN DISTRIBUTED DATABASE SYSTEMS

by

M. R. Stonebraker

E. Wong

Final Technical Report

July 1, 1981 - December 31, 1982

GRANT AFOSR-78-3596

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 83-0512	2. GOVT ACCESSION NO. AD-A129 529	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) CONCURRENT UPDATES AND RETRIEVAL IN DISTRIBUTED DATABASE SYSTEMS		5. TYPE OF REPORT & PERIOD COVERED FINAL, 1 JUL 81-31 DEC 82 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) M.R. Stonebraker and E. Wong		8. CONTRACT OR GRANT NUMBER(s) AFOSR-78-3596
9. PERFORMING ORGANIZATION NAME AND ADDRESS Electronics Research Laboratory University of California Berkeley CA 94720		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE61102F; 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Mathematical & Information Sciences Directorate Air Force Office of Scientific Research Bolling AFB DC 20332		12. REPORT DATE JAN 83 13. NUMBER OF PAGES 11
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

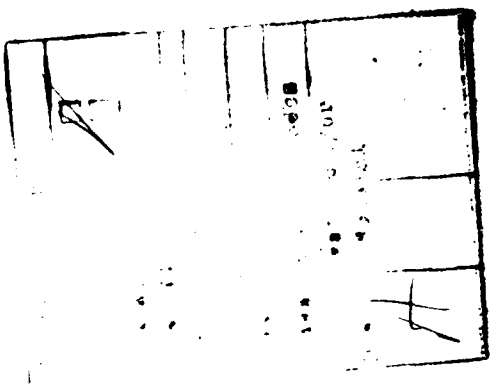
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
At its inception, this project was designed to represent a comprehensive program of research in the field of distributed database management. The problems to be dealt with were to include the three major topics in distributed database: query processing, concurrency control and crash recovery. In addition, the problem of interconnecting heterogeneous databases was also proposed. To a substantial extent, major progress has been achieved in all these areas. In this report a summary of the principal findings is presented.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)



1. Introduction

At its inception, this project was designed to represent a comprehensive program of research in the field of distributed database management. The problems to be dealt with were to include the three major topics in distributed database: query processing, concurrency control and crash recovery. In addition, the problem of interconnecting heterogeneous databases was also proposed.

To a substantial extent, major progress has been achieved in all these areas. In this report a summary of the principal findings is presented.

2. Query Processing

In a distributed DBMS, the database is fragmented, and the fragments distributed, with or without replication. For such systems it is convenient to classify queries into three categories:

(a) Local Queries that can be processed at a single site.

(b) Locally Processable Queries that can be processed at the sites in parallel without any need for intercommunication,

(c) Distributed Queries that do not fall into either of the above classes.

Queries of the first two classes can be processed with no data movement and require no strategies different from query optimization for a centralized database. However, processing of truly distributed queries entails both data movement and strategies that transcend centralized query optimization.

The first general strategy for processing distributed queries was formulated in [WONG77], and a strategy emphasizing parallelism was given in [EPST78]. As a part of the effort for this project, a reformulation that generalizes the above approaches was undertaken [WONG81]. In this formulation, query processing is viewed as an alternating sequence of data movement and local processing. Each operation in the sequence affects in one way or another the data available for processing at each site (collectively referred to as the "materialization" of the database). Query processing, then, can be formulated as a process of "dynamic re-materialization." Viewing the problem this has led to considerable progress on its solution.

Distributed query processing was also studied by Epstein and Stonebraker. In [EPST80] query processing experiments that were performed in a distributed database environment were reported. In this environment several algorithms were compared on the basis of number of bytes moved.

3. Concurrency Control

Locking is a fundamental technique for ensuring data integrity under concurrent accesses. In [RIES79] Ries and Stonebraker studied the problem of choosing the appropriate granularity for locking. The trade-off here is between excessive overhead (small granules) and reduced concurrency (large granules). In [CARE82] granularity hierarchies for locking are considered and several types of concurrency control algorithms are extended to take advantage of such hierarchies.

Multiple-copy consistency is another topic of major interest in distributed concurrency control. In [STON79], possible solutions to this problem are explored. In [CARE83] an abstract model of control

control algorithm is presented.

The model facilitates implementation-independent descriptions of various algorithms, allowing them to be specified in terms of the information that they require, the conditions under which blocking or restarts are called for, and the manner in which requests are processed. The model also facilitates comparisons of the relative storage and CPU overheads of various algorithms based on their descriptions. Results are given for single-site versions of two-phase locking, basic timestamp ordering, and serial validation. Extensions which will allow comparisons of multiple version and distributed algorithms are discussed as well.

4. Crash Recovery

Consistency in a distributed database system is based upon the notion of a transaction, a distributed atomic action. In [SKEE81, SKEE82], Skeen and Stonebraker studied commit protocols for preserving transaction atomicity (and hence consistency) in the presence of failures. They succeeded in:

(1) Introducing a formal framework for reasoning about the crash recovery problem.

(2) Showing fundamental limitations on the fault-tolerance of commit protocols.

(3) Deriving sufficient, and in many cases necessary, properties for a protocol to provide maximum fault-tolerance to various classes of failures.

(4) From the above properties, deriving families of fault-tolerant protocols.

Two failure classes are studied in detail: site failures and network partitioning.

In designing a commit protocol, the primary and overriding objective is to guarantee atomicity; the secondary objective is to maximize availability of the database. Since availability is limited if pending transactions must block (suspend execution) on failures, our focus is on nonblocking protocols.

The formal model introduced is based on nondeterministic finite state automata with failures viewed as a distinguished type of state transition. The model is used both in determining bounds on fault tolerance and in specifying and verifying the protocols summarized below.

Concerning site failures, the Nonblocking Theorem, yielding necessary and sufficient conditions for a commit protocol to be nonblocking was proved. From this result, a family of protocols (the three-phase protocols) was derived. These protocols never require an operational site to block on failures by other sites, even if the transaction coordinator fails.

Concerning site recovery, the nonexistence of nonblocking site recovery was conclusively proved.

Concerning network partitioning, the nonexistence of nonblocking solutions is again proved, and a family of protocols tunable toward maximizing the expected number of nonblocking sites is derived. These protocols are extremely resilient -- resilient when the cause of the failure or even its presence is in doubt.

5. Heterogeneous Databases

If a distributed database system is to integrate existing databases, then a potential problem is heterogeneity. The existing databases may differ in data model and in query language. In [KATZ80, KATZ82, KATZ83] several of the problems associated with heterogeneity were studied.

In [KATZ82] the problem of converting a program expressed in the CODASYL-DML (a procedural language) into a program written in a nonprocedural relational language was studied. The conversion process (dubbed "decompilation") is feasible only under certain circumstances and these are elucidated.

One of the side benefits of the decompilation study was the formulation of a data model (the access path model) that has also found application in physical database design [KATZ83].

6. Implementation

A major effort was undertaken to supplement a distributed version of INGRES. This effort is now complete and a multiple-machine version of INGRES is now operational on a local area network connecting three VAX processors.

REFERENCES

[CARE83a] Carey, M., "Granularity Hierarchies in Conurrency Control," Tech. Mem. M83/1, Electronics Lab., University of California, Berkeley, January, 1983.

[CARE83b] Carey, M., "An Abstract Model of Database Concurrency Control Algorithms," Tech. Mem. M83/6, Electronics Research Lab., University of California, Berkeley, January, 1983.

[EPST78] Epstein, R. S., Stonebraker, M. and Wong, E., "Distributed Query Processing in a Relational Data Base System," Proc. 1978 ACM-SIGMOD Conference on Management of Data, Austin, Texas, May, 1978.

[EPST80] Epstein, R. and Stonebraker, M., "Query Processing in a Distributed Data Base Environment," Proc. 1980 Very Large Data Base Conference, Montreal, Canada, October 1980.

[KATZ79] Katz, R. and Wong, E., "An Access Path Model for Physical Database Design," Proc. ACM-SIGMOD 1980 International Conference on Management of Data, May 1980.

[KATZ82] Katz, R. and Wong, E., "Decompiling CODASYL-DML into Relational Queries," ACM Trans. on Database Systems, Vol. 7, No. 1, March 1982, pp. 1-23.

[KATZ83] Katz, R. and Wong, E., "Resolving Conflicts by Replication in Global Storage Design," ACM Trans. on Database Systems, Vol. 8, No. 1, March, 1983, pp. 110-135.

[RIES79] Ries, D. and Stonebraker, M., "Locking Granularity Revisited," ACM Trans. on Database Systems, Vol. 4, No. 2, June, 1979.

[SKEE81] Skeen, D. and Stonebraker, M., "A Formal Model of Crash Recovery in a Distributed System," 6th Berkeley Workshop on Distributed Data Bases and Computer Networks, Berkeley, CA, February, 1981.

[SKEE82] Skeen, D., "Crash Recovery in a Distributed Database System," Ph.D. Dissertation, University of California, Berkeley, 1982.

[STON79] Stonebraker, M., "Concurrency Control and Consistency of Multiple Copeis of Data in Distributed INGRES," IEEE Trans. on Software Engineering, Vol. 5, No. 2, 1979.

[WONG77] Wong, E., "Retrieving Dispersed Data from SDD-1: A System for Distributed Data Bases," Proc. 2nd Berkeley Workshop on Distributed Data Bases and Computer Networks, Berkeley, CA, May 1977.

[WONG81] Wong, E., "Dynamic Re-Materialization: Processing Distributed Queries Using Redundant Data," Proc. 6th Berkeley Workshop on Distributed Data Bases and Computer Networks, Berkeley, CA, February, 1981.

R.A.'s funded by Grant AFOSR-78-3596

K.P. Birman

D.S. Brunso

M.J. Cary

A. Guttman

S.M. Head

G.W. Mattinger

M.A. Meyer

F.R. Mueller

M. Murphy

R. Probst

J.K. Ranstrom

D.R. Ries

M.D. Skeen

M.A. Whyte

K.C. Wong

D.A. Wood

J.I. Woodfill

Publication Citations

L.A. Rowe, K.P. Birman, "A Local Network Based on the UNIX Operating System," submitted to the IEEE Transactions on Software Engineering.

R.H. Katz, E. Wong, "An Access Path Model for Physical Database Design," published in the 1980 Association for Computing Machinery.

E. Wong, R.H. Katz, "Logical Design and Schema Conversion for Relational and DBTG Databases," published in Entity-Relationship Approach to Systems Analysis and Design, P.P. Chen (ed.) North Holland Publishing Co., 1980.

M. Stonebraker, "Operating System Support for DataBase Management," submitted to Communication of the Association for Computing Machinery.

M. Stonebraker, " Requiem for a Data Base System," submitted to TOD-Summary Guide, September 1980.

R. Epstein, M. Stonebraker, "Analysis of Distributed Data Base Processing Strategies," submitted to the Proceedings of the 1980 Very Large Data Base Conference, Montreal, Canada, October 1980.

D. Skeen, M. Stonebraker, " A Formal Model of Crash REcovery in a Distributed System," submitted to the IEEE Transactions on Software Engineering.

E. Wong, "Dynamic Re-Materialization Processing Distributed Queries Using Redundant Data," presented at the Fifth Berkeley Workshop on Distributed Data Management Computer Networks, Berkeley, California, February 3, 1981.

R.H. Katz, E. Wong, "An Access Path Model for Program Decompilation," submitted to the ACM Transactions on Data Base Systems.

D.R. Ries, M. Stonebraker, "Locking Granularity Revisited," appeared in ACM Transactions on Database Systems, Vol. 4, No. 2, June 1979, pp. 210-227.

M. Stonebraker, "Retrospection on a Data Base System," submitted to the ACM Transactions on Database Systems.

E. Wong, R.H. Katz, "Design Goals for Relational and DBTG Databases," presented at the 1979 SIGMOD Conference in Boston, Massachusetts, May 30 tc June 1, 1979.

Publication Citations

Agrawal, Carey, DeWitt, "Deadlock Detection is Cheap,", submitted to the SIGMOD Record, University of California, E.R.L. Memorandum No. UCB/ERL M83/5, 14 January 1983.

Allman, M. Stonebraker, "Observations on the Evolution of a Software System," University of California, E.R.L. Memorandum No. UCB/ERL M82/59, 2 June 1982.

M. Carey, "Granularity Hierarchies in Concurrency Control," submitted to the 2nd SIGACT-SIGMOD Symposium on Principles on Database Systems, March 1983, Atlanta, GA.

R. Epstein, "Query Processing Techniques for Distributed, Relational Data Base Systems," University of California, E.R.L. Memorandum No. UCB/ERL M80/9, 15 March 1980.

R. Epstein, "Analysis of Distributed Data Base Processing Strategies," published in the Proceedings of the 6th VLDB Conference, Montreal, Canada, October 1982.

A. Guttman, M. Stonebraker, "Using a Relational Database Management System for Computer Aided Design Data," to appear in Data Base Engineering, June 1983.

R.H. Katz, E. Wong, "An Access Path Model for Physical Database Design," published in the Proc. 1980 ACM-SIGMOD International Conference on Management of Data, May 1980.

R.H. Katz, E. Wong, "Decompiling CODASYL DML into Relational Queries," ACM Transactions on Data Base Systems, 1 (March 1982), pp. 1-23.

R.H. Katz, E. Wong, "Logical Design and Schema Conversion for Relational and DBTG Databases," Entity-Relationship Approach to Systems Analysis and Design, P.P. Chen, ed. North-Holland Publishing Co., 1980.

R.H. Katz, E. Wong, "Resolving Conflicts in Global Storage Design through Replication," ACM Transactions on Database Systems, 8 (March 1983) pp. 110-135.

D.R. Ries, M. Stonebraker, "Locking Granularity Revisited," ACM Transactions on Database Systems, Vol. 4, No. 2, June 1979, pp. 210-227.

L.A. Rowe, K.P. Birman, "A Local Network Based on the UNIX Operating Systems, IEEE Transactions on Software Engineering.

M.D. Skeen, "Crash Recovery in a Distributed Database System," to appear in the IEEE Transactions on Software Engineering.

M.D. Skeen, "A Decentralized Termination Protocol," Proc. ACM SIGACT-SIGMOD Conference on Principles of Database Systems, Los Angeles, CA, March 1982.

M.D. Skeen, "Fast Algorithms for VLSI Layout Rule Checking," University of California, E.R.L. Memorandum No. UCB/ERL M81/74, 17 September 1981.

M.D. Skeen, "Nonblocking Commit Protocols," Proc. 1982 ACM-SIGMOD Conference on Management of Data, Orlando, Florida, June 1982.

M. Stonebraker, "Application of Artificial Intelligence Techniques to Database Systems," to appear in Data Base Semantics, edited by Michael Brodic, North Holland Publishers.

M. Stonebraker, H. Stettner, J. Kalash, A. Guttman, N. Lynn, "Document Processing in a Relational Data Base System," too appear in ACM TOOIS.

M. Stonebraker, " Hypothetical Data Bases as Views," Proceedings 1981 ACM-SIGMOD Conference on Management of Data, Ann Arbor Michigan, May 1981.

M. Stonebraker, "Operating System Support for DataBase Management," Communication of the Association for Computing Machinery, June 1981.

M. Stonebraker, J. Woodfill, J. Ranstrom, M. Murphy, J. Kalash, M. Carey, E. Arnold, "Performance Analysis of Distributed Data Base Systems," Data Base Engineering, December 1982.

M. Stonebraker, J. Woodfill, J. Ranstrom, M. Murphy, M. Meyer, E. Allman, "Performance Enhancements to a Relational Data Base System," to appear in ACM Transactions on Database Systems.

M. Stonebraker, "Retrospection on a Data Base System," in TOD-Summary Guide, September 1980.

M. Stonebraker, J. Kalash, "TIMBER: A Sophisticated Relation Browser," Proc. 7th VLDB Conference, Mexico City, September 1983.

J. Woodfill, P. Siegal, J. Ranstrom, M. Meyer, E. Allman, "INGRES Version 7 Reference Manual," University of California, E.R.L. Memorandum No. UCB/ERL M81/61, 27 August 1981.

E. Wong, "Dynamic Re-Materialization Processing Distributed Queries Using Redundant Data," Proc. 6th Berkeley Workshop on Distributed Data Management Computer Networks, Berkeley, California, February 3, 1981.

E. Wong, R.H. Katz, "Logical Design and Schema Conversion for Relational and DBTG Databases," published in Entity-Relationship Approach to Systems Analysis and Design, P.P. Chen, (ed.) North Holland Publishing Co., 1980.